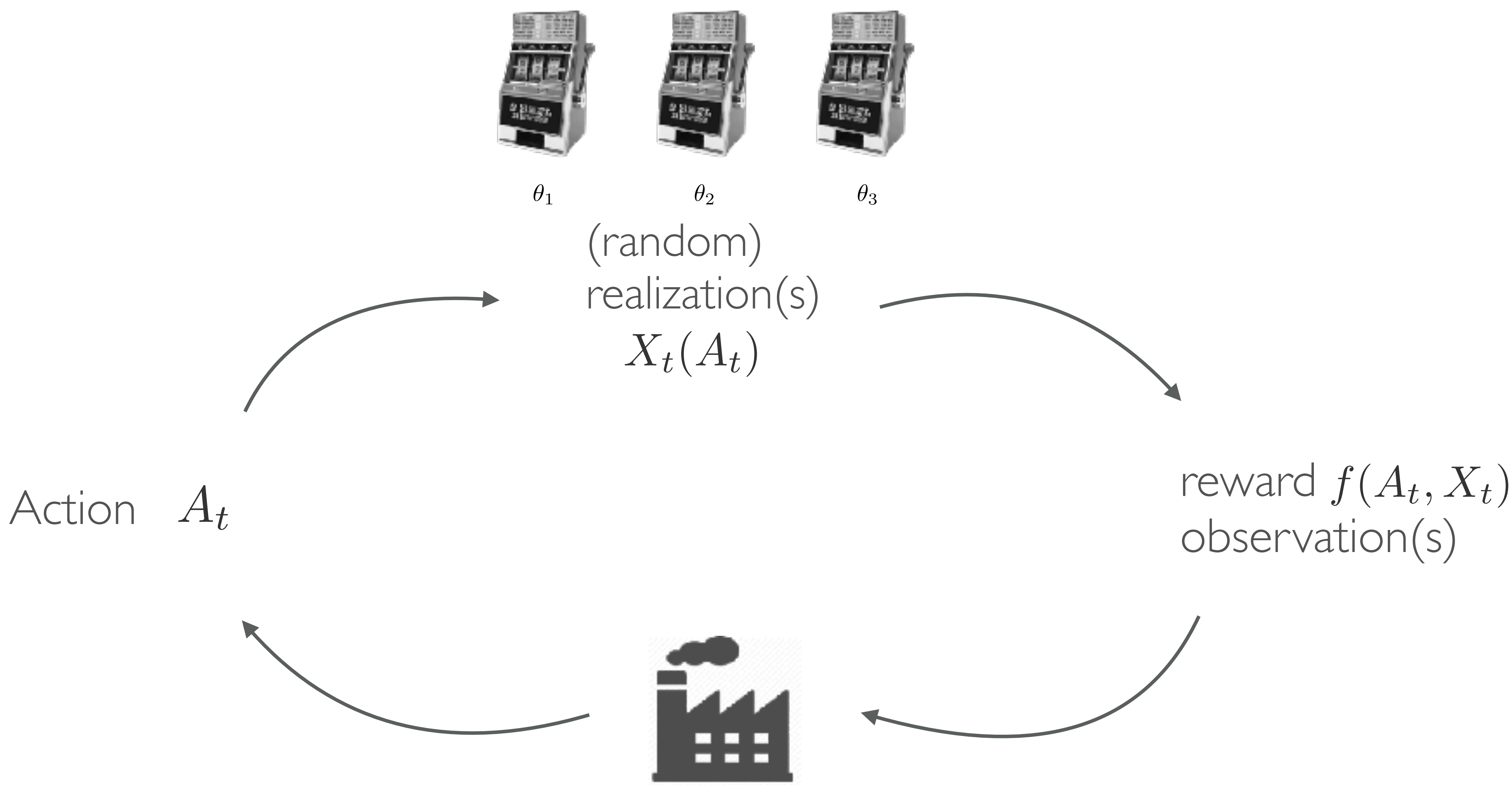
# Delayed Feedback: (Not) Everything Comes to Him Who Waits

**Claire Vernade** Olivier Cappé Vianney Perchet

Amazon CoreAI Berlin
(Télécom ParisTech), CNRS, ENS Paris Saclay, Criteo Research

# Stochastic bandits



# Stochastic bandits

$\theta_1$ $\theta_2$ $\theta_K$

$\longrightarrow X_t(A_t) \sim \mathcal{B}(\theta_{A_t})$

A **policy** (or bandit algorithm) chooses

$$A_t = F(A_1, X_{A_1}, \ldots, A_{t-1}, X_{A_{t-1}})$$

Optimal action: $A^* = \arg\max_i \theta_i$

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{(\theta_{A^*} - \theta_{A_t})}_{\Delta_{A_t}}\right] := \sum_{a \neq A^*} \Delta_a \mathbb{E}[N_a(T)]$$

# Stochastic bandits

Objective: Minimize the regret $R_\theta(T) = \sum_{a \neq A^*} \Delta_a \mathbb{E}[N_a(T)]$

A uniformly efficient policy satisfies for all bandit models $\theta \in \Theta$

$$\forall\, \alpha \in (0,1],\ R_\theta(T) = o(T^\alpha)$$

# Stochastic bandits

Objective: Minimize the regret $R_\theta(T) = \sum_{a \neq A^*} \Delta_a \mathbb{E}[N_a(T)]$

A uniformly efficient policy satisfies for all bandit models $\theta \in \Theta$

$$\forall \alpha \in (0,1],\ R_\theta(T) = o(T^\alpha)$$

Asymptotic lower bound: (Lai & Robbins, 1985)

Every uniformly efficient strategy satisfies

$$\liminf_{T \to \infty} \frac{R(T)}{\log(T)} \geq \sum_{a \neq A^*} \frac{\Delta_a}{d(\theta_a, \theta_{A^*})}$$

where $d(p,q) = p \log\left(\frac{p}{q}\right) + (1-q)\log\left(\frac{1-p}{1-q}\right)$

# Designing Bandit algorithms

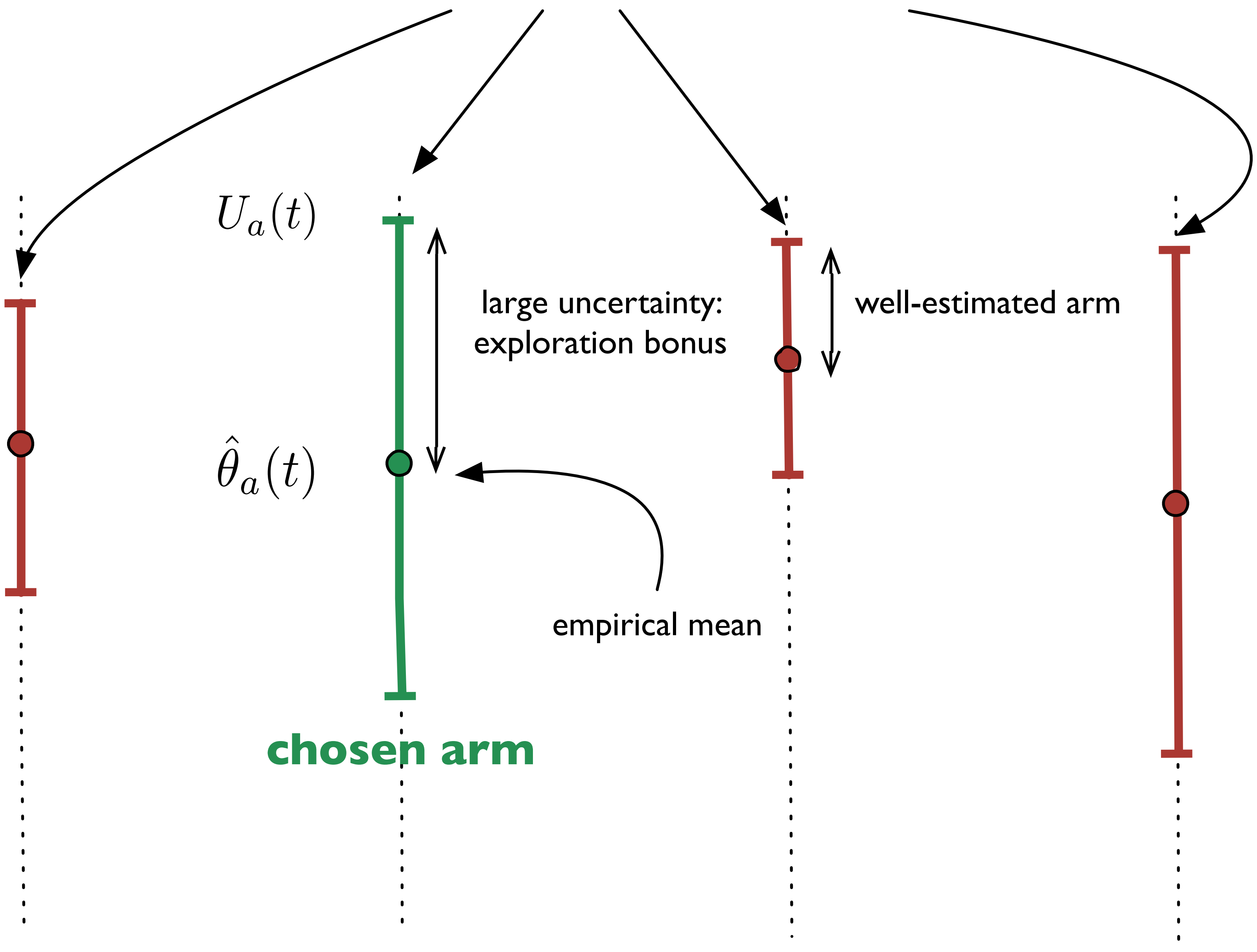
- A bandit algorithm is **asymptotically optimal** if

$$\limsup_{T\to\infty} \frac{R(T)}{\log(T)} \leq \sum_{a\neq A^*} \frac{\Delta_a}{d(\theta_a, \theta_{A^*})}$$

- Asymptotically optimal algorithms for binary bandits:
  - Thompson Sampling (Thompson, 1933) (Kaufmann et al., 2012)
  - KL-UCB (Garivier & Cappé, 2011) (Maillard et al. ,2011)

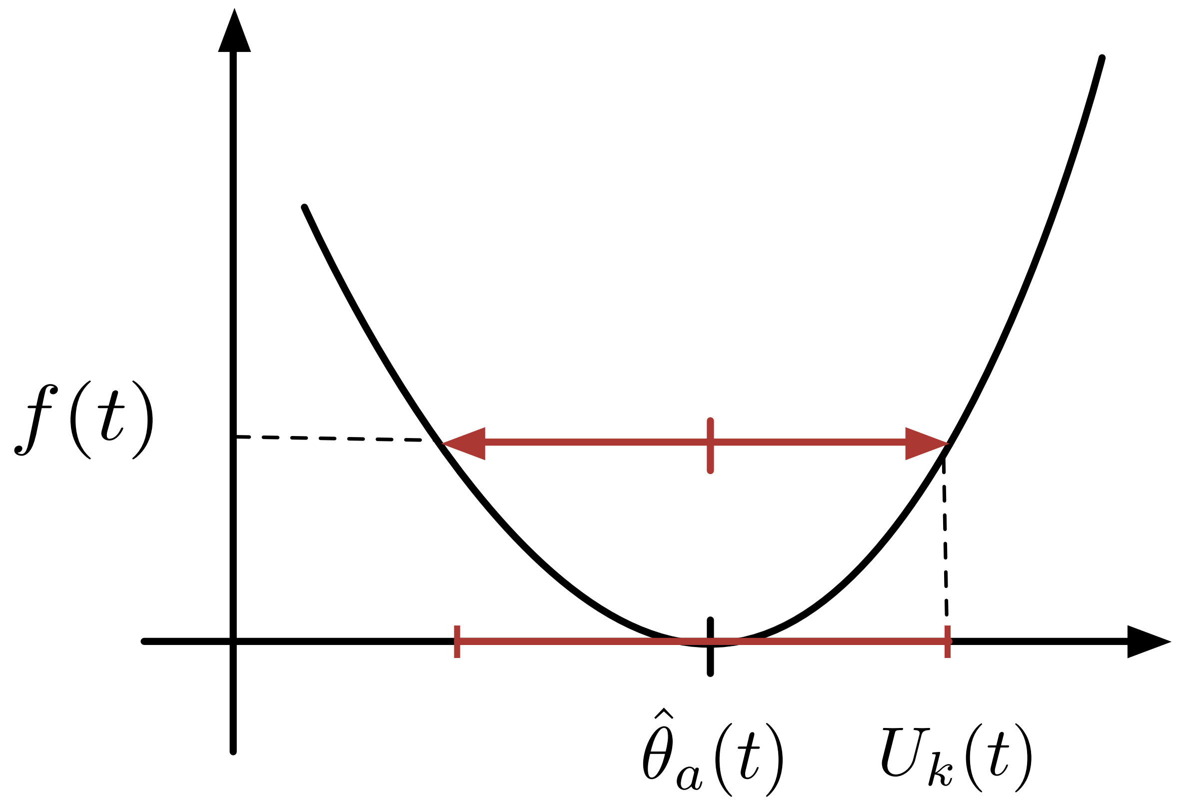
# The optimistic principle: UCB

Upper confidence bounds: $U_a(t) = \hat{\theta}_a(t) + \beta_a(t)$



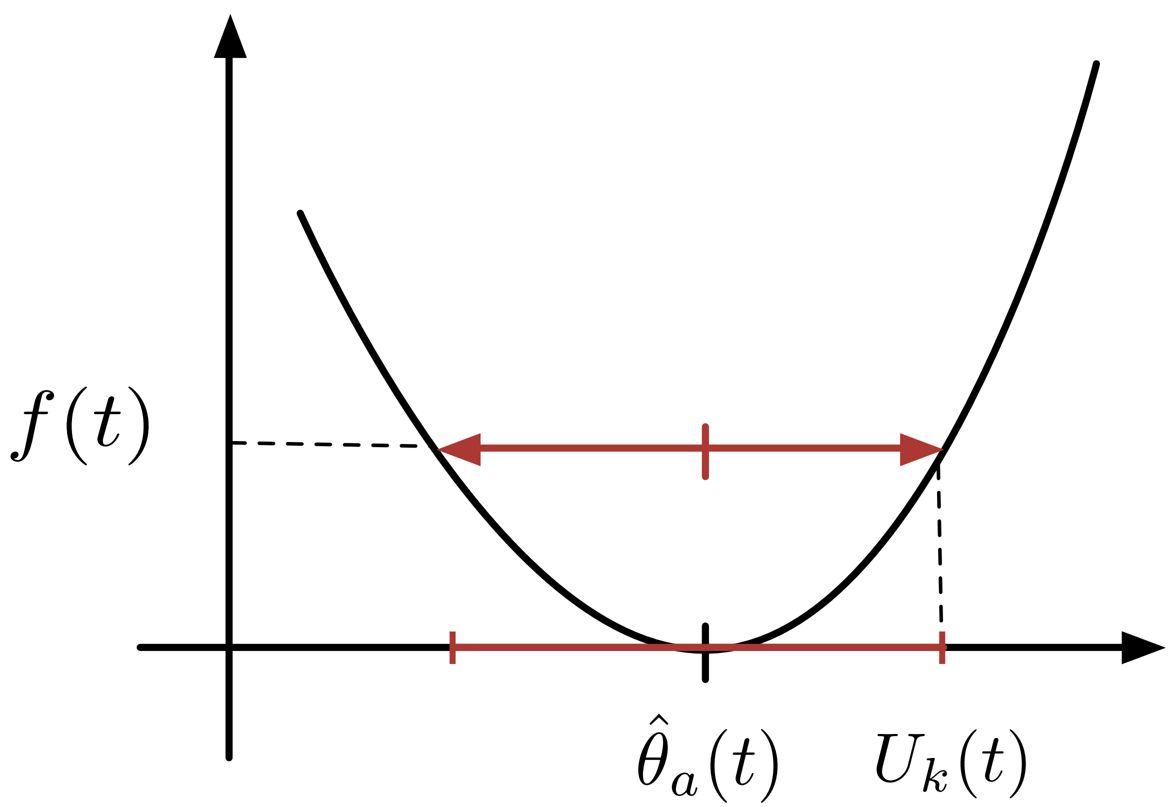
# The optimistic principle: KL-UCB

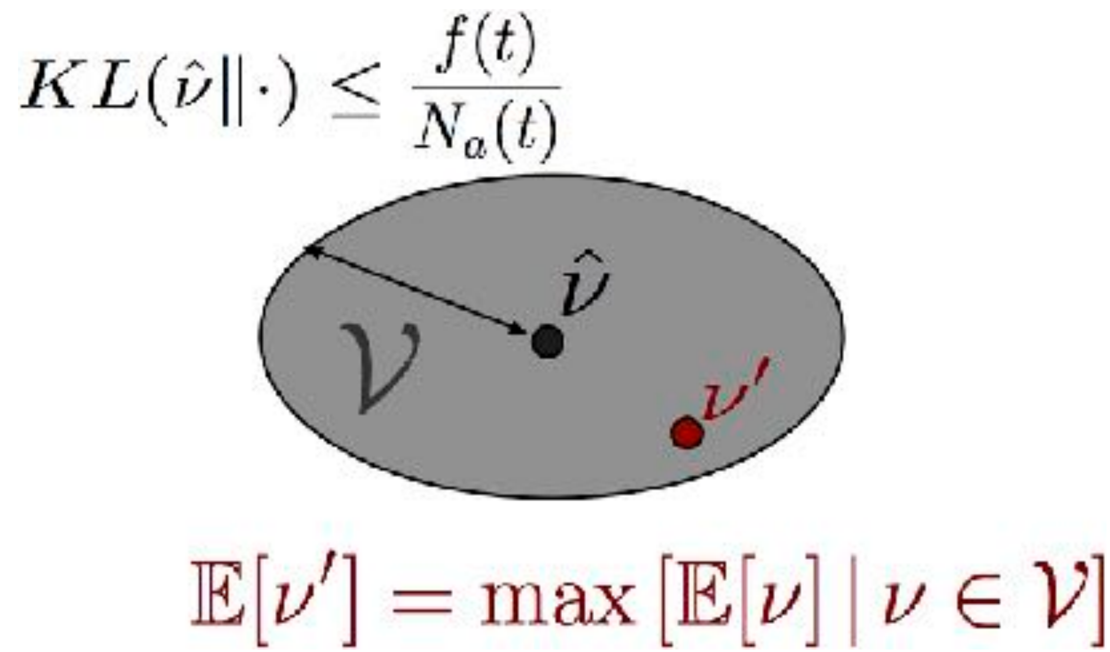
$$U_k(t) = \min_{q \geq \hat{\theta}_k(t)} \left\{ q \,\middle|\, N_k(t) d(\hat{\theta}_k(t), q) \leq f(t) \right\}$$



# The optimistic principle: KL-UCB

$$U_k(t) = \min_{q \geq \hat{\theta}_k(t)} \left\{ q \,\middle|\, N_k(t) d(\hat{\theta}_k(t), q) \leq f(t) \right\}$$



# Summary

Sequential resource allocation problem

Asymptotic lower bound on the regret

(Optimal) algorithms designed on the principle of
Optimism-In-Face-Of-Uncertainty

# Real-world online advertising

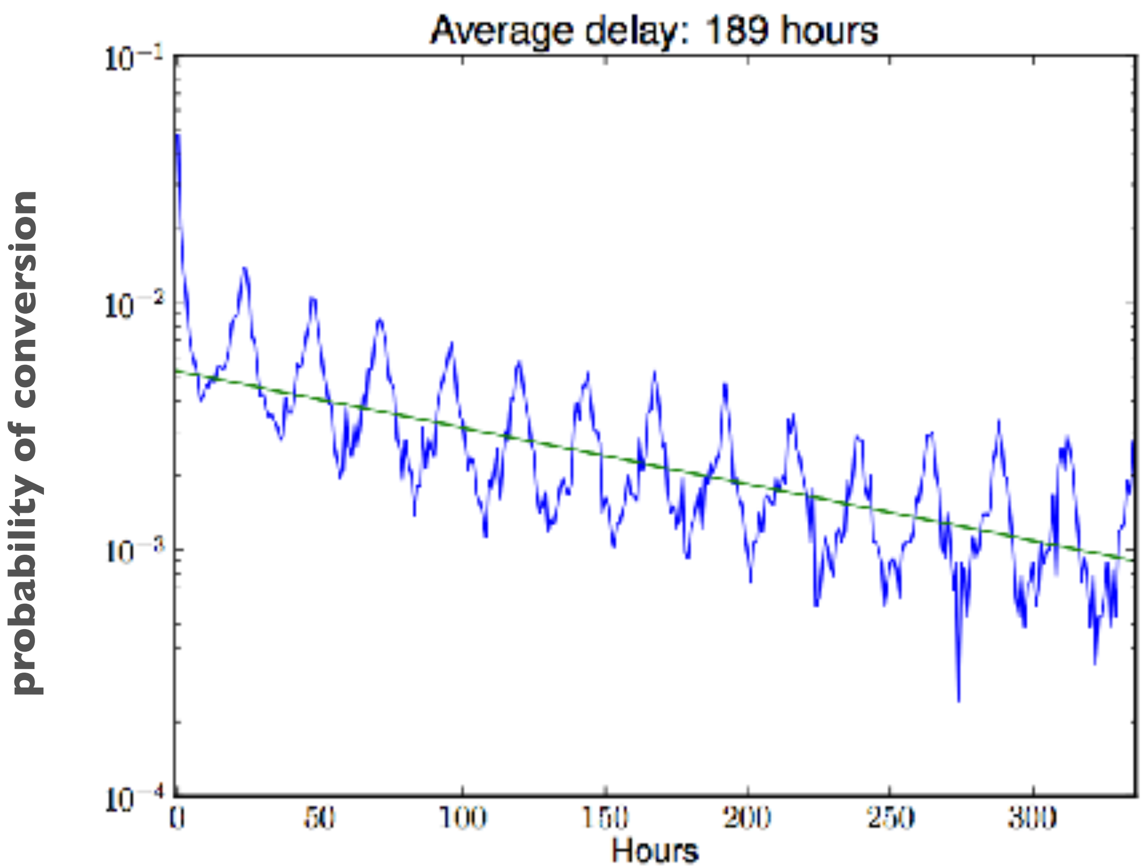



Figure from (Chapelle, 2014)

sequential ad displays

advertising companies expect *conversions:* involved decisions

… which imply delays

displays must be made awaiting feedback

# Real-world online advertising

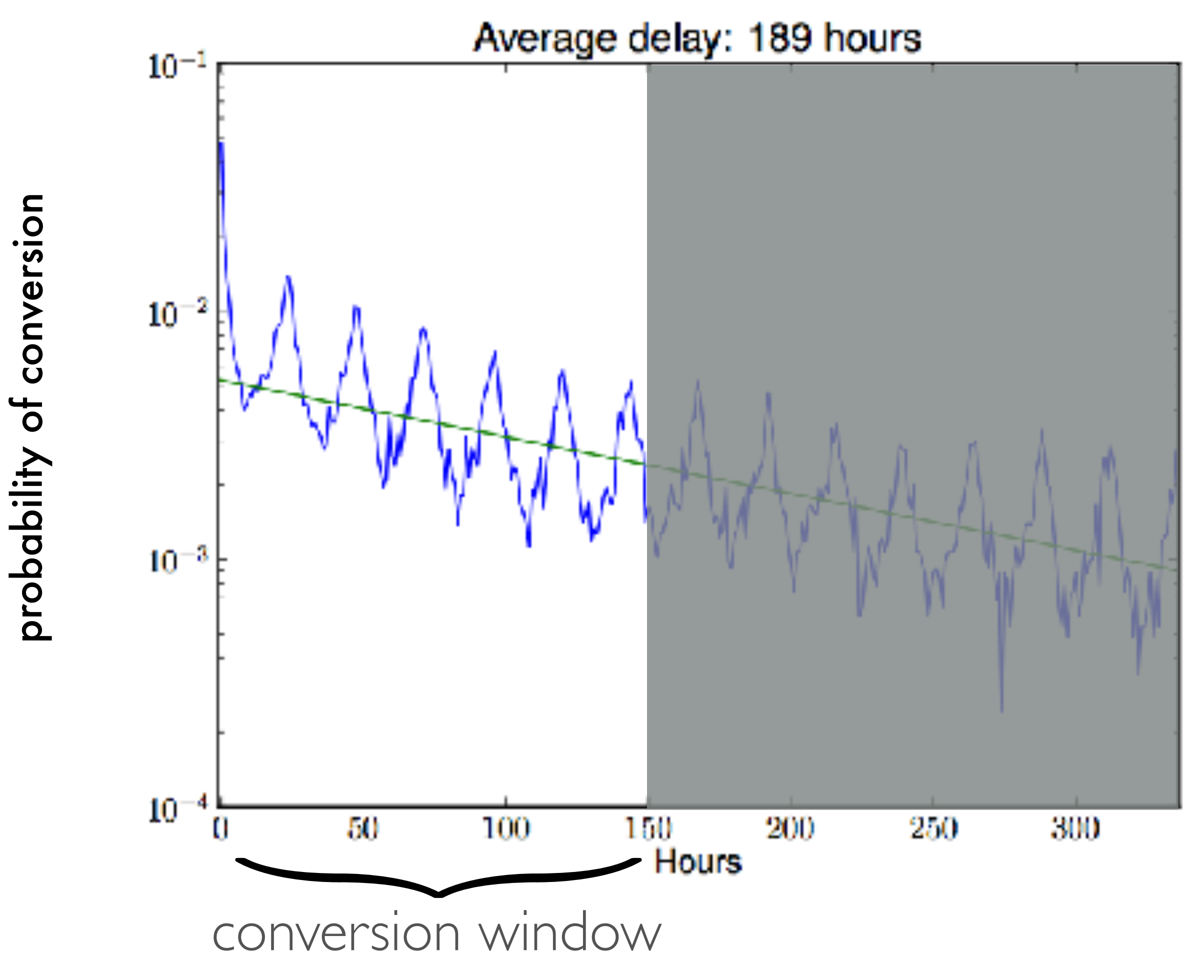



**Figure from (Chapelle, 2014)**

sequential ad displays

advertising companies expect *conversions:* involved decisions

… which imply delays

displays must be made awaiting feedback

# Delayed Conversions

At round $t_1$…

$t_1$

$$A_{t_1} \in \{1, \dots, K\}$$

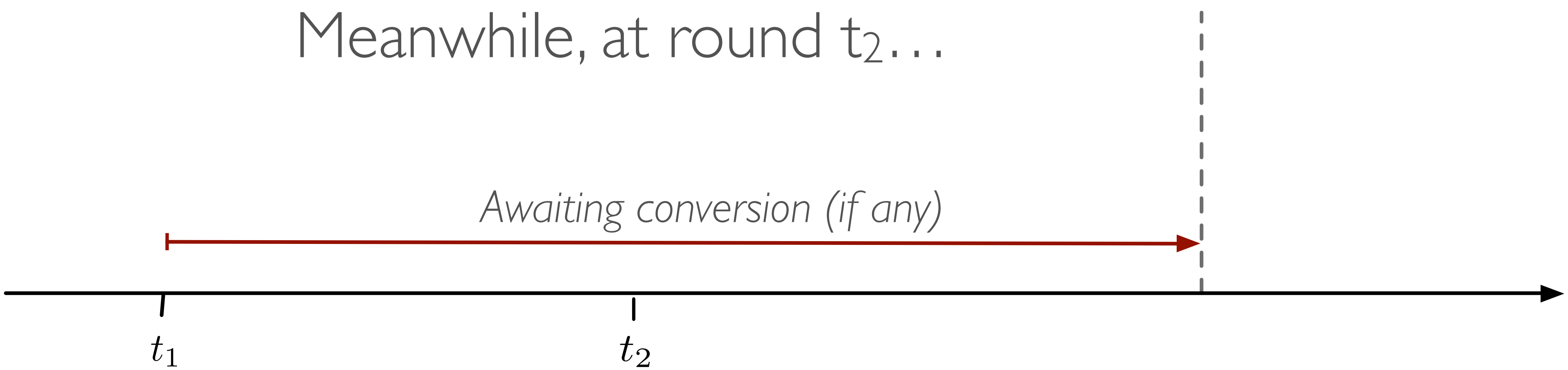
$$\begin{cases} C_{t_1} \sim \mathcal{B}(\theta_{A_{t_1}}) \\ D_{t_1} \sim \text{CDF } \tau \end{cases}$$

Binary conversion indicator

Delay before disclosure

# Delayed Conversions

Meanwhile, at round $t_2$…




$A_{t_1} \in \{1, \ldots, K\}$

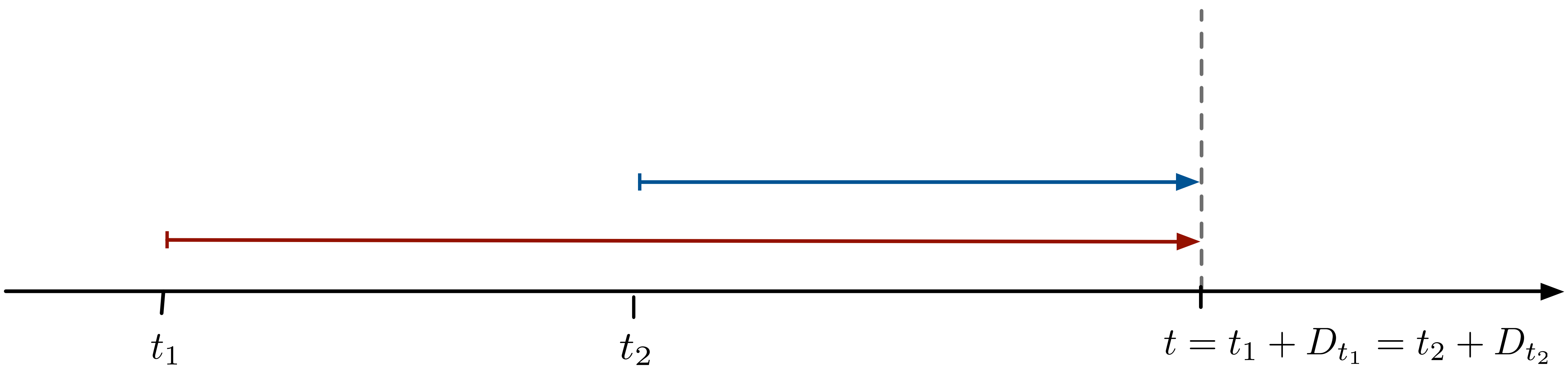
$$\begin{cases} C_{t_1} \sim \mathcal{B}(\theta_{A_{t_1}}) \\ D_{t_1} \sim \text{CDF } \tau \end{cases}$$

$A_{t_2}$

$$\begin{cases} C_{t_2} \\ D_{t_2} \end{cases}$$

# Delayed Conversions

At round t, conversions are disclosed…



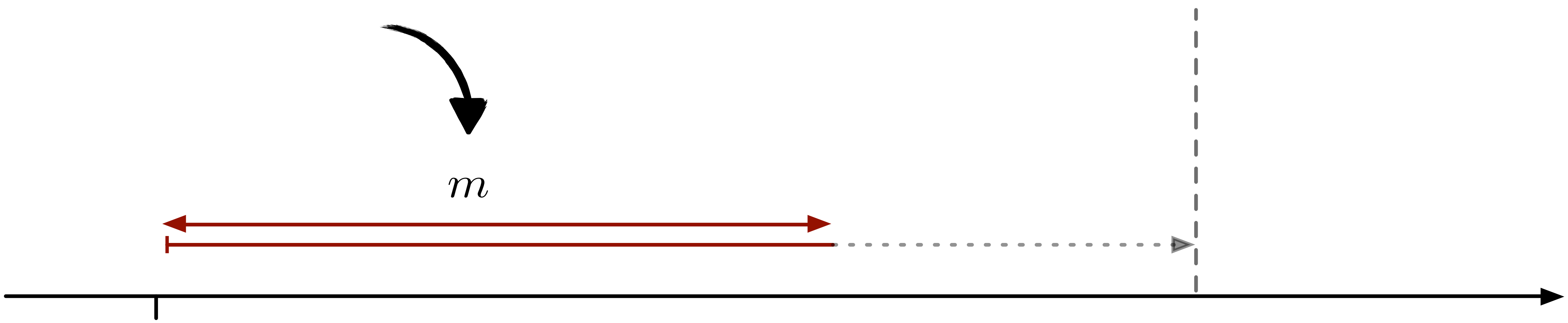

$A_{t_1} \in \{1, \dots, K\}$ $\quad A_{t_2}$ $\quad A_t \dots$

$$\begin{cases} C_{t_1} \sim \mathcal{B}(\theta_{A_{t_1}}) \\ D_{t_1} \sim \text{CDF } \tau \end{cases} \qquad \begin{cases} C_{t_2} \\ D_{t_2} \end{cases} \qquad Y_t = C_{t_1} + C_{t_2}$$

Reward at round t:
sum of all due conversions

# Thresholded feedback

Waiting time capped to m time steps!

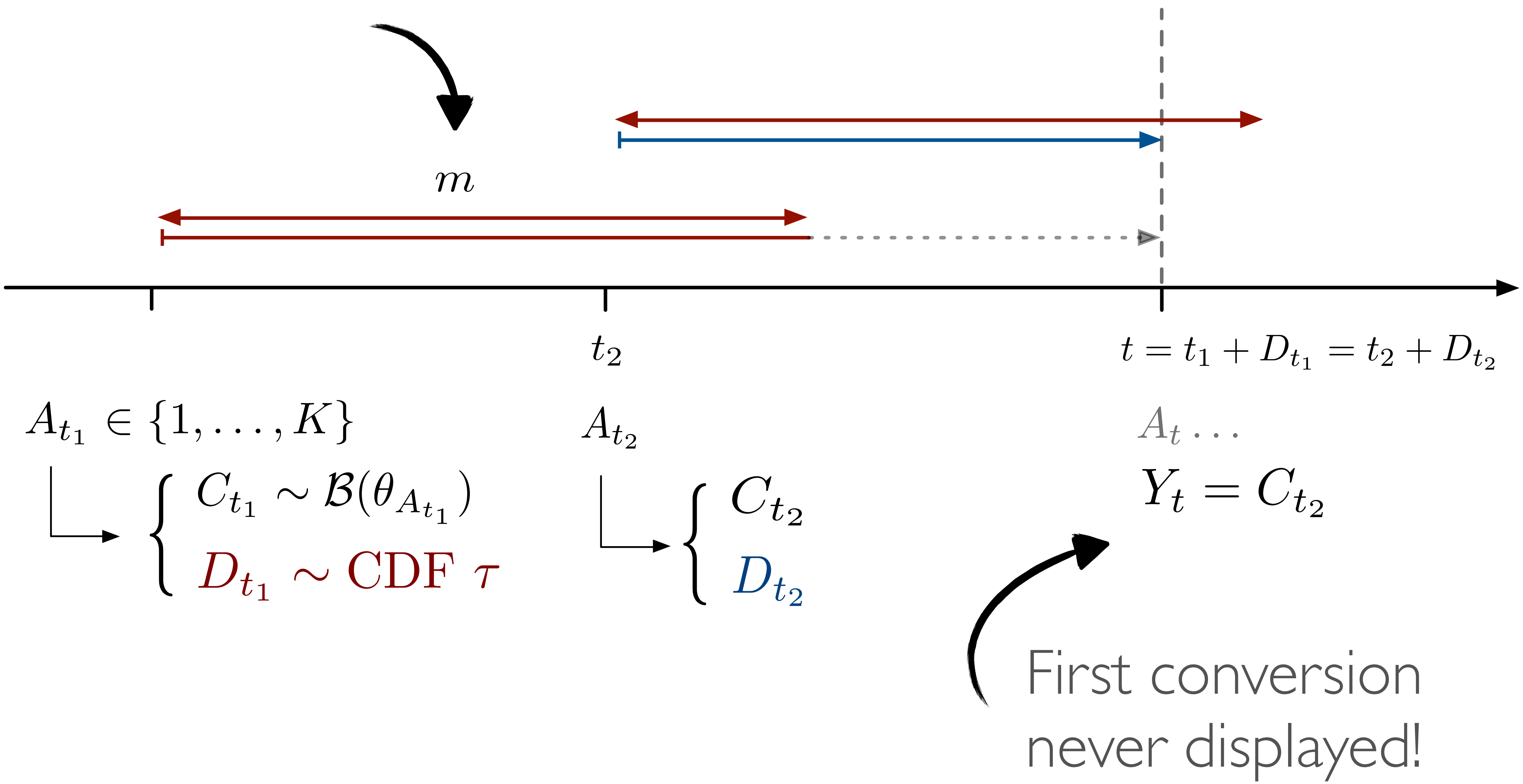



$$A_{t_1} \in \{1, \ldots, K\}$$

$$\begin{cases} C_{t_1} \sim \mathcal{B}(\theta_{A_{t_1}}) \\ D_{t_1} \sim \text{CDF } \tau \end{cases}$$

# Thresholded feedback

Waiting time capped to m time steps!



# Regret minimization

The regret of an algorithm is defined by

$$R(T) = \sum_{s=1}^{T} \mathbb{E}\left[\theta_{a^*} - \theta_{A_s}\right] \tau_{T-s}$$

# Regret minimization

The regret of an algorithm is defined by

$$R(T) = \sum_{s=1}^{T} \mathbb{E}\left[\theta_{a^*} - \theta_{A_s}\right] \tau_{T-s}$$

losses are weighted
according to their delay

# Regret minimization

The regret of an algorithm is defined by

$$R(T) = \sum_{s=1}^{T-m} (\theta_{a^*} - \mathbb{E}[\theta_{A_s}])\tau_m + \sum_{s=T-m+1}^{T} (\theta_{a^*} - \mathbb{E}[\theta_{A_s}])\tau_{T-s}$$

old pulls:
only a proportion
of the rewards
will be disclosed

most recent
pulls

# Lower bound on the regret

Assumption: the expectation of the delays is bounded

Theorem 1: (Uncensored case)
For uniformly efficient policy, we prove that

$$\liminf_{T \to \infty} \frac{R(T)}{\log(T)} \geq \sum_{a \neq A^*} \frac{\Delta_a}{d(\theta_a, \theta_{A^*})}$$

We retrieve the Lai & Robbins' bound:
the uncensored delayed problem is not harder !

# Lower bound on the regret

Assumption: the expectation of the delays is bounded

Theorem II: (Censored case)

For *uniformly efficient policy*, we prove that

$$\liminf_{T \to \infty} \frac{R(T)}{\log T} \geq \sum_{k \neq *} \frac{\tau_m(\theta^* - \theta_k)}{d(\tau_m \theta_k, \tau_m \theta^*)}$$

# Lower bound on the regret

Assumption: the expectation of the delays is bounded

Theorem II: (Censored case)
For *uniformly efficient policy*, we prove that

$$\liminf_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{k\neq *} \frac{\tau_m(\theta^* - \theta_k)}{d(\tau_m\theta_k, \tau_m\theta^*)} \geq \sum_{k\neq k^*} \frac{(\theta^* - \theta_k)}{d(\theta_k, \theta^*)}.$$

Lai and Robbins' bound

The problem in the censored case is harder.

# Optimistic policies

Unbiased estimators under delayed feedback:

1. Corrected counts of observations:

$$\tilde{N}_k(t) = \underbrace{\sum_{s=1}^{t-m} \mathbf{1}\{A_s = k\}\tau_m}_{\substack{\text{old pulls: a proportion} \\ \text{has never been disclosed} \\ \text{due to censoring}}} + \underbrace{\sum_{s=t-m+1}^{t-1} \mathbf{1}\{A_s = k\}\tau_{t-s}}_{\substack{\text{recent pulls: weighted} \\ \text{according to current delay}}}$$

# Optimistic policies

Unbiased estimators under delayed feedback:

2. Estimator of $\theta$ :

Cumulated rewards

$$\hat{\theta}_k(t) = \frac{\sum_{s=1}^{t-1} C_s \mathbf{1}\{A_t = k\}}{\tilde{N}_k(t)}$$

Delay-corrected
number of pulls

# Optimistic policies

Unbiased estimators under delayed feedback:

3. Optimistic index, D-UCB algorithm:

Exploration rate

$$U_k(t) = \hat{\theta}_k(t) + \sqrt{\frac{N_k(t)}{\tilde{N}_k(t)}}\sqrt{\frac{\beta_\epsilon(t)}{2\tilde{N}_k(t)}},$$

Correction factor due to missing observations

# Optimistic policies

Unbiased estimators under delayed feedback:

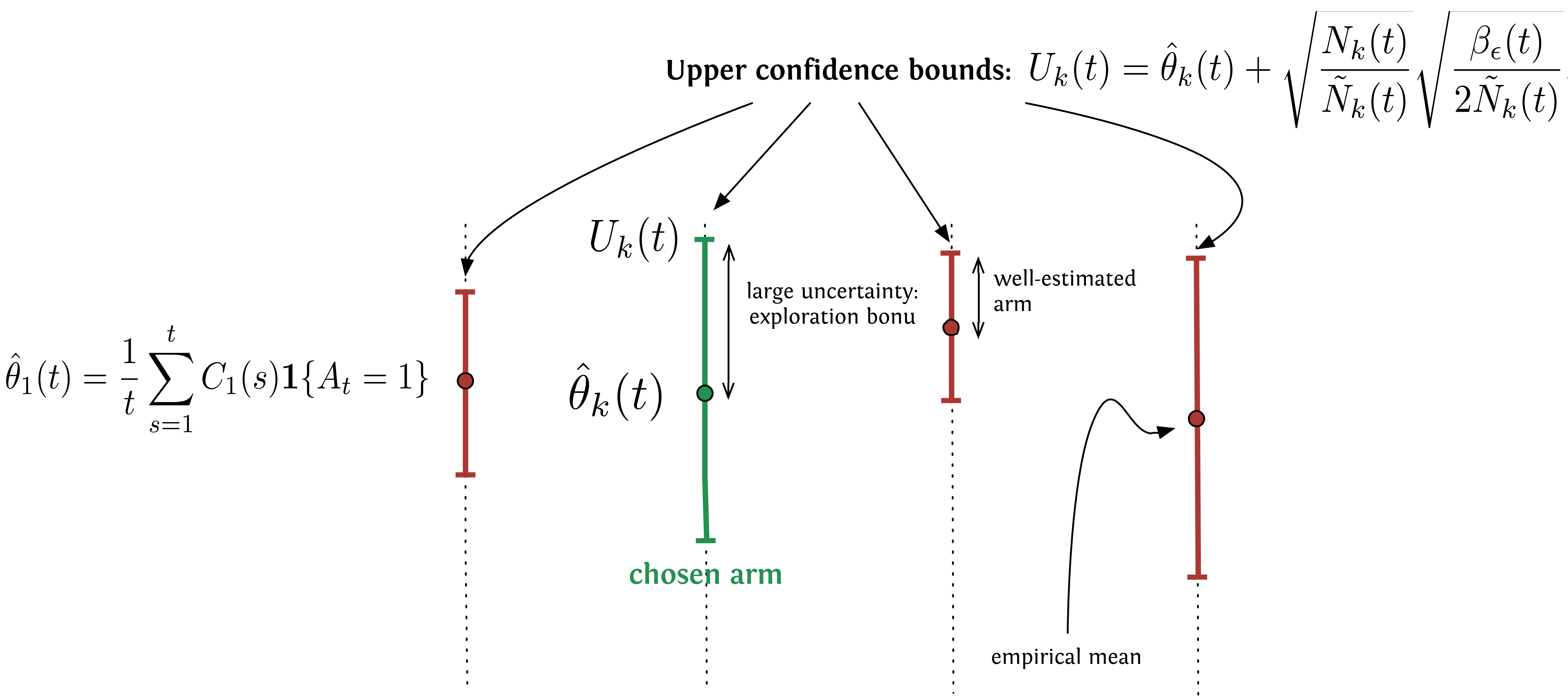
4. Optimistic index, D-KL-UCB algorithm:

$$U_k^{\mathrm{KL}}(t) = \max_{q \geq \hat{\theta}_k(t)} \left\{ q \,|\, \tilde{N}_k(t) d_P(\hat{\theta}_k(t), q) \leq (1+\epsilon)\log(t) \right\}$$

Corrected count of pulls: expected nb of observations

KL-divergence between **Poisson** distributions !

# Optimistic policies



# Regret Analysis

The regret of D-UCB is bounded by

$$R(T) \leq (1+\epsilon)\log(T)\sum_{k\neq *}\frac{1}{2\tau_m\Delta_k} + o_{\epsilon,m}(\log(T)).$$

The regret of D-KL-UCB is bounded by

$$R(T) \leq (1+\epsilon)\log(T)\sum_{k\neq *}\frac{\tau_m\Delta_k}{d_P(\tau_m\theta_k, \tau_m\theta^*)} + o_{m,\epsilon}(\log(T))$$

# Experiments

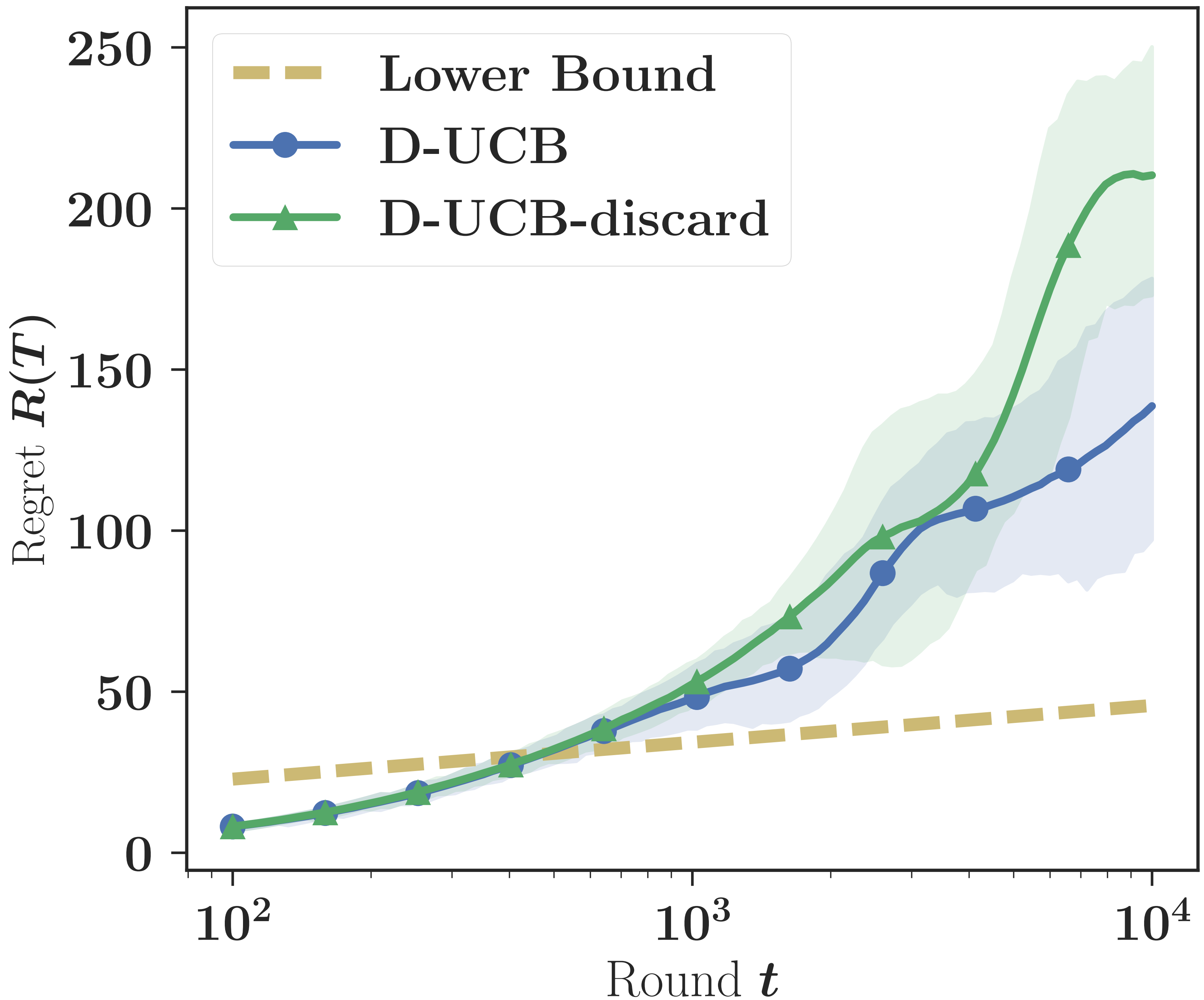
- We compare our algorithms D-UCB and KL-UCB to the immediate baseline DISCARD:
  - Store rewards and actions within the conversion window,
  - Only use the data available at t-m to make prediction at t+1.
  - Should be optimal asymptotically, but in practice …

# Experiments

3 arms: 0.7, 0.5, 0.3.

Expected delay = 500

m=1000

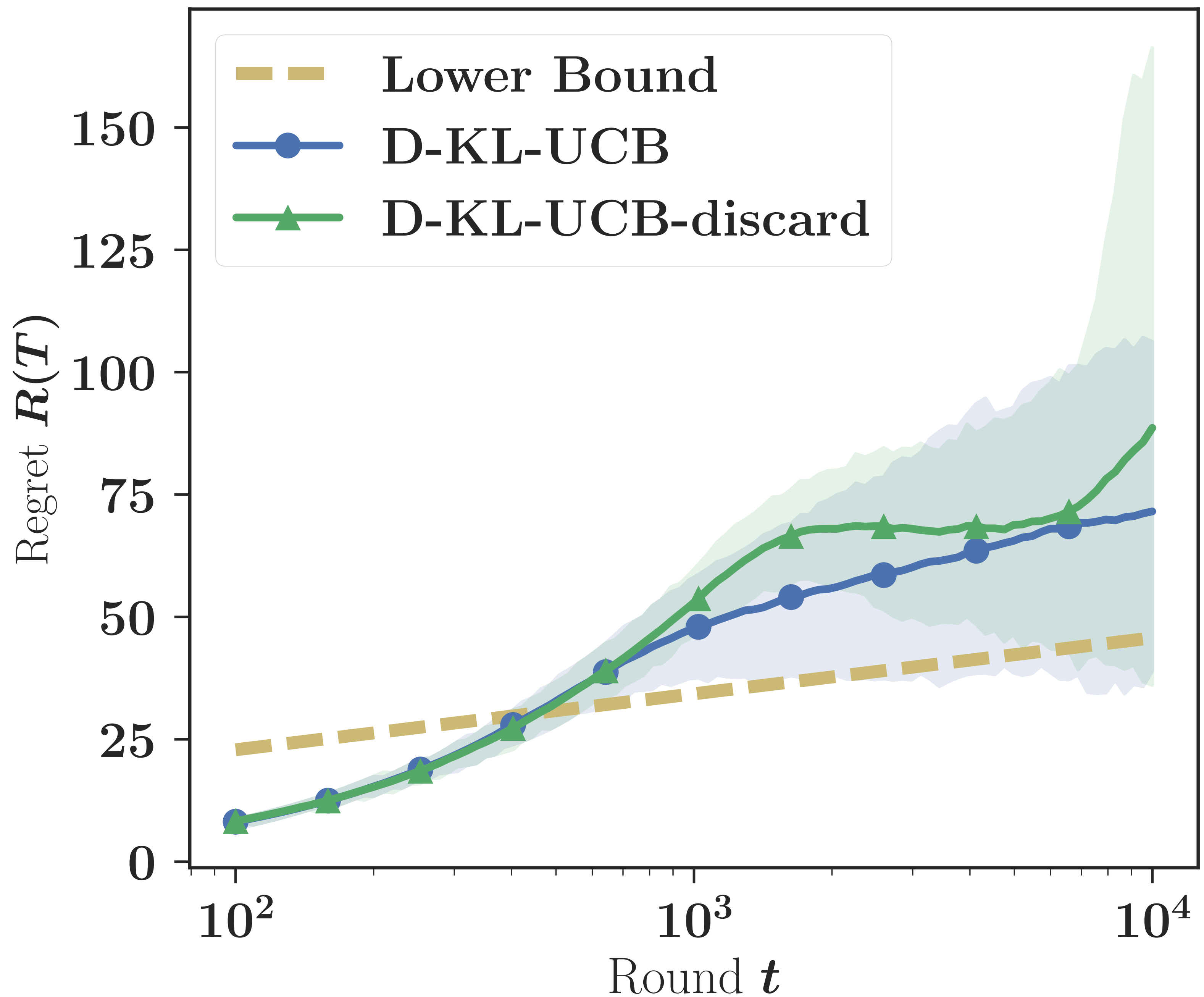


# Experiments

3 arms: 0.7, 0.5, 0.3.

Expected delay = 500

m=1000



# Conclusion

New bandit model under censored delayed feedback

Problem-dependent lower bound

(asymptotically optimal) optimistic algorithms

# Conclusion

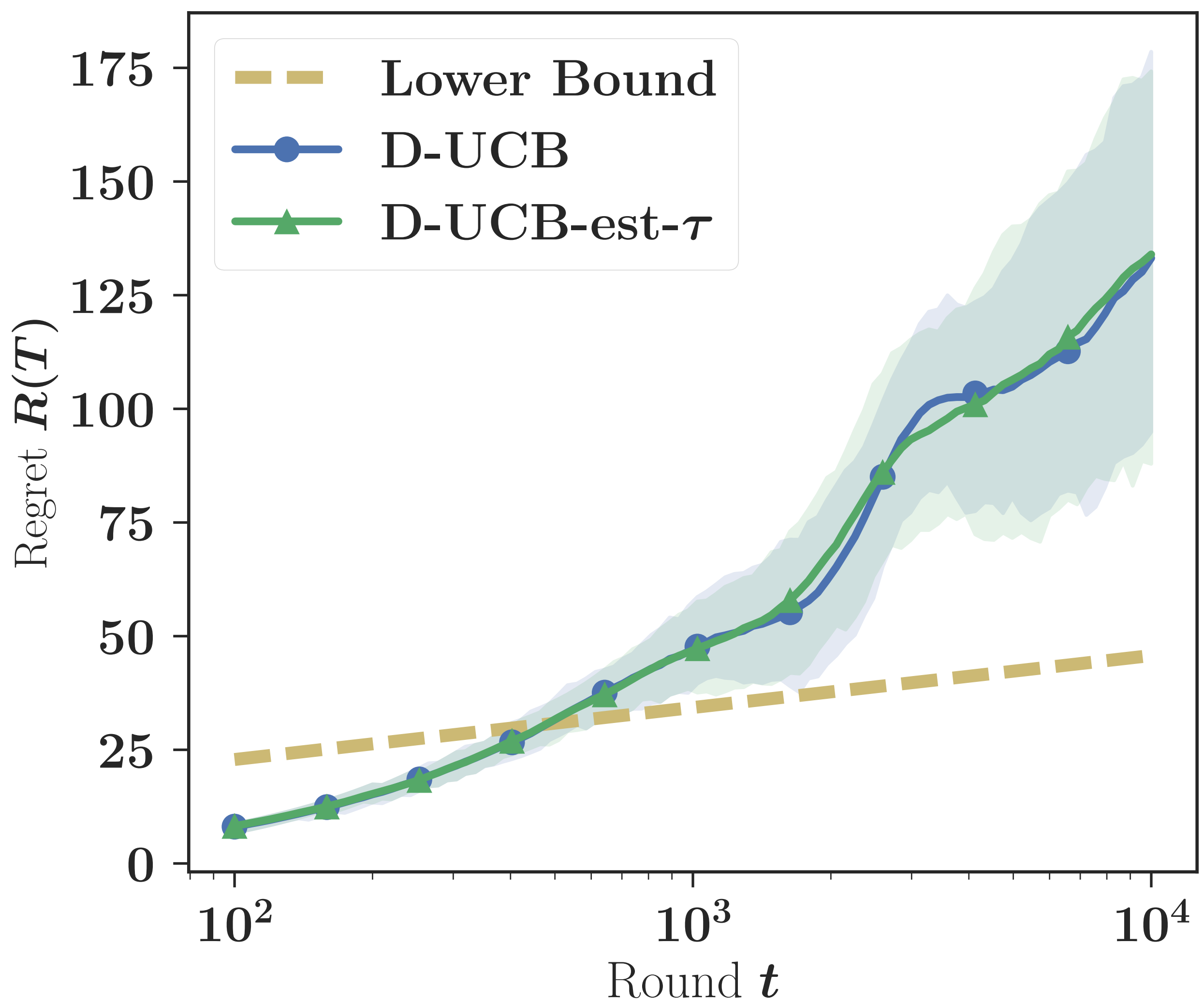
New bandit model under censored delayed feedback

Problem-dependent lower bound

(asymptotically optimal) optimistic algorithms

Requires prior knowledge of the delay distribution

# Conclusion




Open problem : analysis of this heuristic that seems to work empirically